1 of 1

AD 603 876

ON A THEOREM OF DOOB

T. E. Harris

P - 139 ✓ gnd

17 April 1950

Approved for OTS release

10 p
hc - 1.00
mf - 0.50

The RAND Corporation
· SANTA MONICA · CALIFORNIA

# ON A THEOREM OF DOOB

T. E. Harris

**SUMMARY**

This note gives a justification for the interchange of limiting processes required in Doob's "heuristic" approach to the Kolmogorov limiting distribution of the maximum deviation between a theoretical and an empirical distribution function.

. . . . . . . . . . . . . .

A number of writers have recently treated the Kolmogorov-Smirnov limiting distributions from the point of view of stochastic processes. ( 1 , 2 , and 3 .) For example, let

$$D_n = \sup_{-\infty < x < \infty} \sqrt{n}\,\left| F(x) - F_n(x) \right|$$

where $F(x)$ is an arbitrary continuous cumulative distribution and the random variable $F_n(x)$ is the sample cumulative: $nF_n(x)$ is the number of values which are $\leq x$ in a sample of $n$ from a population described by $F(x)$.

Since the limiting distribution of $D_n$ is the same for any continuous $F$, it is sufficient to let $F(t) = t$, $0 \leq t \leq 1$, and to consider

$$u_n(t) = \sqrt{n}\,(F_n(t) - t), \qquad 0 \leq t \leq 1,$$

where the $n$ sample values are picked according to the uniform distribution on $(0,1)$. For any fixed set of values $t_1, \ldots, t_k$, the joint distribution of $u_n(t_1), \ldots, u_n(t_k)$ approaches that of $y(t_1), \ldots, y(t_k)$ as $n \to \infty$, where

$y(t)$ is the Gaussian process defined by

$$y(t) = x(t) - tx(1),$$

$x(t)$ being the Wiener process. This is the heuristic guide to the fact shown by Doob [2] that the limiting distribution of $D_n$ is the same as the distribution of

$$D = \sup_{0 \leq t \leq 1} |y(t)| .$$

It appears worthwhile to give a simple and rigorous justification for the transition from $D_n$ to $D$. Such questions come up frequently, and are also of some theoretical interest in connection with "Monte Carlo" procedures where continuous stochastic processes are approximated with random walks.

Let

$$\theta(z) = P(D \leq z)$$

$$\theta_n(z) = P(D_n \leq z) = P\left(\sup_{0 \leq t \leq 1} u_n(t) \leq z\right) .$$

We wish to show that for any $z$,

$$(1) \qquad \lim_{n \to \infty} \theta_n(z) = \theta(z) .$$

The desired result will follow from Theorem 1.

<u>Theorem 1. Let $a$ and $b$ be arbitrary positive numbers. Then $n_0$ and $\Delta_0 > 0$ can be determined so that for all $n \geq n_0$</u>

$$P\left\{\max_{0 \le t_1 \le t_2 \le t_1 + \Delta_0 \le 1} \left|u_n(t_2) - u_n(t_1)\right| > a\right\} < b.$$

We make use of an idea going back to W. K. Clifford 1866; see Moran [4]. This is the fact that if $T_1, \ldots, T_{n+1}$ are independent random variables each having the density $ce^{-ct}$ for any $c > 0$, the quantities $T_j/(T_1 + \cdots + T_{n+1})$ are jointly distributed like the n+1 intervals which are obtained when n points are picked uniformly and independently at random on the interval (0,1). In other words let $G_n(t)$ be a Poisson process with rate n; i.e., the probability of a jump of + 1 between t and t + dt is $ndt + O(dt)^2$. Let $L_n$ be the time when the (n+1)st jump occurs. Then the two stochastic processes $F_n(t)$ and $G_n(L_n t)$ are equivalent for t in (0,1). Now for large n, $L_n$ converges stochastically to 1, and thus to prove theorem 1 we consider first $G_n(t)$.

Let

$$v_n(t) = \sqrt{n}\left[G_n(t)/n - t\right],$$

$$H_{a,n}(\Delta) = P\left[\max_{0 \le t \le \Delta} \left|v_n(t)\right| \ge a\right].$$

Lemma. $$H_{a,n}(\Delta) \le \frac{4}{\sqrt{2\pi}} \int_{a/\sqrt{2\Delta}}^{\infty} e^{-\frac{1}{2}y^2}\,dy$$

for $n \ge n(a, \Delta)$.

This lemma could be strengthened by general methods used by Erdos and Kac [5], but the following simple derivation is sufficient here. Let T be the smallest t-value for which $v_n(t) \ge a$, let $K_{an}(T)$ be the cumulative

distribution of T and let

$$Q_{an}(t) = P \left[ v_n(t) \geq a \right] .$$

Then

$$(2) \qquad Q_{an}(2\Delta) \geq \int_0^{2\Delta} Q_{an}(2\Delta - T)\, dK_{an}(T)$$

$$\geq \int_0^{\Delta} Q_{an}(2\Delta - T)\, dK_{an}(T)$$

The first $\geq$ sign in (2) appears because $v_n(t)$ is a discontinuous differential process; for a continuous differential process equality would hold. Now

$$Q_{an}(2\Delta) \rightarrow \frac{1}{\sqrt{2\pi}} \int_{a/\sqrt{2\Delta}}^{\infty} e^{-\frac{1}{2}y^2}\, dy$$

as $n \rightarrow \infty$, and $Q_{an}(\sigma) \rightarrow \frac{1}{2}$ uniformly in $\sigma$ for $\Delta \leq \sigma \leq 2\Delta$. Hence (2) implies that

$$\frac{2}{\sqrt{2\pi}} \int_{a/\sqrt{2\Delta}}^{\infty} e^{-\frac{1}{2}y^2}\, dy \geq K_{an}(\Delta)$$

for $n \geq n(a, \Delta)$. A similar argument applies to

$$K^{*}_{an}(\Delta) = P\left[\min_{0 \le t \le \Delta} v_n(t) \le -a\right],$$

and the lemma follows from the obvious fact that

$$H_{an}(\Delta) \le K_{an}(\Delta) + K^{*}_{an}(\Delta).$$

A standard type of argument now tells us that theorem 1 holds if $u_n(t)$ is replaced by $v_n(t)$. For we may divide the interval $(0,1)$ into $k$ equal parts of length $\frac{1}{k}$. The probability that on none of these intervals does $v_n(t)$ differ from its initial value on that interval by more than $\frac{a}{4}$ is

$$\left[1 - H_{\frac{a}{4},n}\left(\frac{1}{k}\right)\right]^{k} = P_k .$$

We may then choose $\Delta_0 = \frac{1}{k_0}$ small enough so that

$$\left[1 - \frac{4}{2\pi}\int_{a/(4\sqrt{2\Delta})}^{\infty} e^{-\frac{1}{2}y^2}\,dy\right]^{k} \ge 1 - b$$

for $\frac{1}{\Delta} = k \ge k_0$. By virtue of lemma 1, we can then choose $n_0$ so that

$$\left[1 - H_{\frac{a}{4},n}(\Delta_0)\right]^{k_0} \ge 1 - b$$

for $n \ge n_0$. Theorem 1 now holds for $v_n(t)$, with the quantities $a$, $b$, $\Delta_0$, and $n_0$ just chosen, since with probability $\ge 1 - b$, $v_n(t)$ has no oscillations

$\geq \frac{a}{2}$ in any of the intervals $(\frac{j-1}{k}, \frac{j}{k})$ and thus no oscillations $\geq a$ in any interval of length $\Delta_o$.

Now

$$(3) \qquad v_n(L_n t) = \sqrt{n}\left(\frac{G_n(L_n t)}{n} - L_n t\right) = u_n(t) + t\,\xi_n$$

where the random variable $\xi_n$,

$$\xi_n = \sqrt{n}\,(1-L_n)$$

is asymptotically normal with zero mean and unit variance.

It is easily seen that theorem 1 must also hold on (0,1) for the random function $v_n(L_n t)$, which is produced from $v_n(t)$ by a random magnification of the t-scale; for an arbitrary $c > 0$ we can make the probability arbitrarily high, by taking n large enough, that the magnification $L_n$ is between 1-c and 1+c. Theorem 1 likewise holds for the random function $t\,\xi_n$; it must therefore hold for the difference

$$u_n(t) = v_n(L_n t) - t\,\xi_n.$$

The following scheme now gives (1).

Define

$$S_n = \sup_{0 \le t \le 1} \left| u_n(t) \right|$$

$$S_n^k = \sup_{0 \le j \le k} \left| u_n\left(\frac{j}{k}\right) \right|$$

$$\theta_n(z) = P(S_n \le z)$$

$$\theta_n^k(z) = P(S_n^k \le z)$$

$$\psi(z) = P\left[ \sup_{0 \le t \le 1} \left| y(t) \right| \le z \right]$$

$$\psi^k(z) = P\left[ \sup_{0 \le j \le k} \left| y\left(\frac{j}{k}\right) \right| \le z \right] .$$

We use the known result that $\psi(z)$ is continuous. Let $z \ge 0$ and $\alpha > 0$ be given. Pick $\epsilon > 0$ so that $\left| h \right| \le \epsilon$ implies $\left| \psi(z) - \psi(z + h) \right| < \alpha/4$. Then, using theorem 1 and observing that

$$\theta_n(z) \ge P(S_n \le z,\ S_n^k \le z - \epsilon) =$$

$$P(S_n^k \le z - \epsilon) - P(S_n^k \le z - \epsilon,\ S_n > z)$$

we may, by theorem 1, pick $n'$ and $k'$ so that $n > n'$, $k > k'$, implies

$$\theta_n^k(z - \epsilon) \le \theta_n(z) + \alpha/4.$$

Next pick $k'' \geq k'$ so that $k \geq k''$ implies

$$\left| \psi(z) - \psi^{k}(z) \right| < \alpha/4,$$

$$\left| \psi(z-\varepsilon) - \psi^{k}(z-\varepsilon) \right| < \alpha/4.$$

Then choose $n'' \geq n'$ so that $n \geq n''$ implies

$$\left| \psi^{k''}(z) - \theta_n^{k''}(z) \right| < \alpha/4,$$

$$\left| \psi^{k''}(z-\varepsilon) - \theta_n^{k''}(z-\varepsilon) \right| < \alpha/4.$$

For all $n \geq n''$ we then have

$$(4) \qquad \psi(z) \leq \psi(z-\varepsilon) + \alpha/4 \leq \psi^{k''}(z-\varepsilon) + 2\alpha/4 \leq$$

$$\theta_n^{k''}(z-\varepsilon) + 3\alpha/4 \leq \theta_n(z) + \alpha,$$

$$(5) \qquad \psi(z) \geq \psi^{k''}(z) - \alpha/4 \geq \theta_n^{k''}(z) - \alpha/2 \geq \theta_n(z) - \alpha/2.$$

From (4) and (5) it follows that for all $n \geq n''$

$$\psi(z) - \alpha \leq \theta_n(z) \leq \psi(z) + \alpha/2$$

and this establishes (1).

## REFERENCES

[1]. T. W. Anderson and D. A. Darling, "Some Statistical Problems Connected with Stochastic Processes," to appear.

[2]. J. L. Doob, "Heuristic Approach to the Kolmogorov-Smirnov Theorems," Annals of Math. Stat., 20 (1949), 393-403.

[3]. M. Kac, "On Deviations between Theoretical and Empirical Distributions," Proc. Nat. Acad. of Sciences, 35 (1949), 252-257.

[4]. P. A. P. Moran, "The Random Division of an Interval," Journal of the Royal Statistical Society Supplement, 9 (1947), 92-98.

[5]. P. Erdos and M. Kac, "On Certain Limit Theorems of the Theory of Probability," BAMS, 52 (1946), 292-302.

TKH:jo